El consumo de información: una primera aproximación al concepto

Gemma Martínez

Citación recomendada: Gemma Martínez. *El consumo de información: una primera aproximación al concepto* [en linea]. "Hipertext.net", núm. 1, 2003. <http://www.hipertext.net> [Consulta: 12 feb. 2007]. .

1. Introducción

En el ámbito de la documentación [1] a menudo se hace referencia a la problemática de la recuperación de la información, a la importancia de la definición de perfiles de usuario o de la difusión selectiva de la información. Un factor común a todas ellas es lo que se denomina a menudo como una necesidad informativa, que normalmente se concibe de un modo apriorístico.

El objeto del presente artículo es ofrecer una nueva perspectiva del análisis de las necesidades informativas de una persona introduciendo el concepto de *consumo de información* . Se trata de analizar a posteriori el modo en que un usuario, en tanto que persona que trabaja con información y documentos, lleva a cabo su actividad a partir del rastro que dicha actividad deja, por ejemplo, en el ordenador personal que utiliza.

A partir de la exposición de una hipótesis de trabajo concreta pasaremos a enumerar y describir los datos que se consideran significativos de este consumo informativo, los instrumentos empleados para la localización de dichos datos, y las posibles aplicaciones técnicas de trabajo. A continuación, expondremos algunos ejemplos significativos de prototipos de agentes de información que se basan en análisis del consumo informativo. En último lugar, intentaremos hacer una reflexión sobre las posibilidades que este planteamiento de trabajo a posteriori ofrece para poder llevar a cabo una difusión de la información lo más ajustada a las necesidades informativas reales de un usuario y/o de un colectivo informativo.

2. Acerca del consumo de información

Como apuntamos anteriormente, en el ámbito de la documentación estamos habituados a términos como perfil de usuario (*user profile*), definición de perfil de usuario (*user profiling*), necesidad informativa (*information need*). El término que introducimos en el presente artículo, consumo de información, se basa en conceptos no tan habituales como uso de información (*information use*), comportamiento informativo (*information behaviour*), o información "just in time" (*just-in-time information*). Entendemos consumo de información como la acción de buscar y recuperar datos e información con la finalidad de dar respuesta a una necesidad o interés informativo. No podemos hablar, por lo tanto, de consumo de información sin vincularlo estrechamente a la definición de perfil de usuario. El consumo informativo nos aporta datos sobre las acciones y habilidades informativas del usuario, y en consecuencia, nos lo define a partir de una perspectiva nueva desde el punto de vista de la documentación.

¿Qué significa la incorporación de datos sobre la actividad informativa real? ¿Cómo nos puede ser útil conocer el

consumo real de información de un usuario, conocer sus habilidades informativas, para completar su perfil a priori?

La hipótesis de trabajo de la que partimos es la siguiente: si conocemos el uso informativo de los trabajadores de un organismo o empresa, siempre en relación a las tareas inherentes a su puesto, el servicio de documentación de dicho organismo o empresa podrá ofrecer un servicio de difusión de la información más personalizado y preciso, dado que la definición de perfiles de usuario se hará de manera más exacta.

En este contexto, por lo tanto, consumidor de información es el usuario que accede a los recursos de información, gratuitos y de pago, que el organismo en el que trabaja pone a su disposición con el objetivo de dar respuesta a necesidades informativas que, en nuestro caso, serán laborales. Es decir, excluimos necesidades personales y/o de ocio, dado que entendemos que deben ser cubiertas fuera del marco del organismo o empresa. Así pues, estamos restringiendo este análisis a usuarios que utilizan un ordenador / terminal para llevar a cabo las tareas asignadas, y, concretamente, nos interesamos por la información que en este ordenador quedará registrada de su uso de los programas de trabajo Office y del programa de navegación utilizado (Netscape o Explorer), dado que los datos registrados son altamente significativos de su consumo informativo.

El objetivo final de este análisis es la creación de un agente de información que pueda ofrecer a cada trabajador la información necesaria para realizar sus tareas con sólo iniciar su sesión de trabajo en el ordenador / terminal de su empresa. De este modo, la difusión de la información se llevaría a cabo de manera justo a tiempo, minimizando el tiempo de búsqueda informativa, rendibilizando al máximo las fuentes de información disponibles y siendo lo más pertinente y eficaz posible.

Pasamos a enumerar y describir brevemente la tipología de datos que nos hablan de este consumo informativo y su localización, diferenciando entre Office e Internet.

3. Los datos sobre consumo de información: descripción, clasificación y localización

La primera etapa de nuestro proceso de trabajo se ha basado en la localización, identificación y clasificación de los datos de nuestra actividad en Microsoft Office y en Internet que quedaban recogidos en el ordenador de trabajo, a partir de la observación directa de los instrumentos de localización como son el Explorador de Windows, o los propios programas de navegación y de bibliografía sobre experiencias sobre análisis de comportamiento informativo. Nos han sido de gran utilidad los estudios de los autores que a continuación citamos concretamente por la tipología de datos en los que han basado sus análisis: Davison (2002) trabaja sobre predicción de comportamientos de navegación en la web a partir de los contenidos de los recursos consultados ; Damiani [et al.] (2001) realiza un similar estudio al anterior pero basado en el análisis del historial de navegación del usuario ; Hicks (1999) propone el uso los metadatos como base para la customización de servicios de información.

Para establecer una clasificación de los datos seleccionados como significativos del consumo de información nos hemos basado, principalmente, en la categorización, muy acertada en nuestra opinión, de los datos que formulan Adomavicius & Tuzhilin (2001). Estos autores hacen una diferenciación entre lo que denominan datos factuales y datos transaccionales de un usuario. Mientras que los datos factuales nos dan información sobre la identidad del usuario, nombre, sexo, fecha de nacimiento, ingresos anuales, etc., los datos transaccionales recogen la actividad del usuario. Así, si escogiéramos analizar una compra por Internet, dicha transacción nos daría información sobre el producto escogido, su precio, el tipo de tarjeta de crédito usada, entre otros datos. La elaboración de un perfil de usuario sencilla se basa únicamente en datos factuales, es decir, en lo que el usuario es dado su sexo, su edad, las tareas asignadas, etc.

Esta diferenciación entre datos factuales y transaccionales permite diferenciar dos tipos de definición de perfil de usuario: vertical y horizontal. La definición de un perfil de usuario vertical viene dada por los datos factuales del usuario, mientras que el perfil horizontal surge del análisis de los datos transaccionales del consumo informativo del usuario. Si aplicamos estas definiciones a un grupo de usuario concreto, el perfil vertical respondería a la definición individual de cada uno de ellos a partir de sus datos factuales y responde a la visualización de cada individuo como un ente propio y claramente definido básicamente a partir de las tareas asignadas. La definición de perfiles de usuario horizontales, en cambio, surge del análisis transversal de los datos transaccionales de dicho grupo. Es a partir de este análisis que aparecen comportamientos informativos similares y/o dispares en función de necesidades informativas, compartidas o no, entre los diferentes usuarios del grupo, de ahí que el concepto de horizontal responda a la visualización de los resultados del análisis transversal de diferentes perfiles verticales.

Basándonos pues en la categorización de Adomavicius & Tuzhilin (2001) hemos establecido dos bloques para clasificar los datos sobre consumo: por un lado formal / contenido y, por el otro, factual / transaccional. Un dato de tipología formal / contenido hace referencia propiamente al tipo de recurso o documento. Así, por ejemplo, una URL indica el tipo de conexión que se establece (http, ftp, telnet) o su dominio nos puede dar información sobre su finalidad (.net, .edu, .org., etc.) o su país de origen (.es, .ca, .uk, etc.). Los metadatos de una página web proporcionan información formal y especialmente, según los casos, sobre su contenido, a partir de la enumeración de palabras clave.

Un dato de tipología factual / transaccional da información sobre el usuario, quién es y qué hace. Por ejemplo la propiedad *Autor* en el conjunto de propiedades de un informe redactado con el programa Word nos aporta a menudo el nombre y apellido/s de la persona que lo ha realizado. La relación de documentos Office abiertos, creados y/o modificados que queda recogida en la carpeta /Reciente (Windows 2000) informa sobre la actividad llevada a cabo por dicha persona durante los últimos días.

Sintetizamos en las siguientes tablas los datos recogidos de nuestra actividad, por un lado, en Internet y, por el otro, en Microsoft Office. En dichas tablas se recogen la tipología y localización de cada dato. Su ordenación, de mayor a menor importancia, responde a criterios que exponemos a continuación.

Tabla 1. Actividad en Internet

DATO	**TIPOLOGÍA**	**LOCALIZACIÓN**
Contenido de la página consultada (textual e hipertextual)	Contenido/factual	Memoria caché
Palabras clave utilizadas en los buscadores consultados	Formal/transaccional	Memoria caché
Metadatos de la página consultada	Contenido/factual	Memoria caché
Dirección de Internet consultada (URL)	Formal/factual	Memoria caché
Páginas web seleccionadas (Favoritos)	Formal/contenido/transaccional	Bookmarks / Favoritos del navegador
Relación de les páginas consultadas: Hace dos semanas Cada día de la semana en curso incluido el día de hoy Por sede web, haciendo una relación de las páginas dependientes consultadas Por número de visitas más alto (de mayor a menor) en general Por número de visitas más alto en el día de hoy	Formal/contenido/transaccional	Historial de navegación
Cookies residentes en el ordenador	Formal/transaccional	Memoria caché
Tiempo (fecha de caducidad, fecha de la última modificación [caché], fecha del último acceso, fecha de la última comprobación)	Formal/factual	Memoria caché
Nombre del fichero	Formal/contenido/factual	Memoria caché

Tipo de fichero	Formal/factual	Memoria caché
Medida del fichero	Formal/factual	Memoria caché

La ponderación de los datos referentes a la actividad en Internet se justifica a partir de los estudios anteriormente citados de Davison (2002) y Hicks (1999) sobre predicción de comportamientos de usuario y costumización de servicios de información. Nuestra conclusión es que los datos que más información pueden aportar sobre el consumo de información con un grado de fiabilidad más alto son los propios contenidos de los documentos y recursos consultados y sus elementos descriptivos, es decir, los metadatos. De aquí que estos dos datos ocupen la primera y la tercera posición respectivamente. Las palabras clave recogidas en la formulación de la búsqueda son especialmente relevantes en la medida en que nos dicen el modo en que el usuario ha explicitado su necesidad informativa. Así pues, los contenidos y los metadatos de los recursos consultados nos confirmarían la pertinencia y relevancia de los recursos obtenidos. No menos importantes son las páginas web seleccionadas como útiles, interesantes, etc., así como el historial de búsquedas. El resto de datos son de tipo formal pero no por ello desechables.

Es remarcable la importancia de la memoria caché del programa de navegación como instrumento de trabajo para la localización de los datos descritos.

Tabla 2. Actividad en Microsoft Office (*) A/P/G de ahora en adelante

DATO	**TIPOLOGÍA**	**LOCALIZACIÓN**
Relación de documentos creados, modificados o editados recientemente	Formal/contenido/transaccional	Explorador de Windows 2000 (C:Documents and settings *Carpeta usuario* Reciente)
Contenido textual y hipertextual del documento	Contenido/factual	Editor de textos (Word)
Nombre del documento	Formal/contenido/factual	Archivo/Propiedades/General (*)
Autor	Formal/contenido/factual	A/P/G
Tema (Asunto)	Contenido/factual	A/P/G
Palabras clave	Contenido/factual	A/P/G
Categoría	Contenido/factual	A/P/G
Base del hipervínculo	Formal/contenido/factual	A/P/G
Tiempo de edición	Formal/factual	A/P/G

Fecha de la última consulta	Formal/factual	A/P/G
Fecha de creación	Formal/factual	A/P/G
Fecha de modificación	Formal/factual	A/P/G
Ubicación	Formal/factual	A/P/G
Tamaño del fichero	Formal/factual	A/P/G
Comentarios	Contenido/factual	A/P/G
Número de revisiones hechas	Formal/factual	A/P/G

En cuanto a la aplicación de los parámetros basados en los estudios de Davison (2002) y Hicks (1999), si en el caso de la actividad en la web el contenido de los recursos y documentos consultados quedan residentes en el ordenador de trabajo gracias a la memoria caché, en el caso de la actividad en Office esta información queda reflejada, de manera formal, en carpetas que el sistema operativo crea de manera automática (concretamente Windows 2000). Es por este motivo que dicha relación, que puede ser considerada dato y instrumento de localización a la vez, aparezca encabezando la tabla. El resto de datos hacen referencia, una vez más, al contenido del documento y a los datos que lo describen. Conviene remarcar el gran número de datos que se recogen en las propiedades del propio documento y que pueden ser significativas para la identificación de los contenidos y posterior seguimiento del consumo informativo, como son el tema (asunto) o las palabras clave.

4. Técnicas de análisis

Dado que uno de nuestros objetivos es la obtención de perfiles de usuario horizontales a partir del análisis de los datos transaccionales resultantes del consumo informativo es probable que el volumen de datos a analizar sea muy elevado, por lo que lo más adecuado será, sin duda, el uso de técnicas de análisis automatizadas que nos permitan llevar a cabo esta tarea de manera rápida y fiable.

Por esta razón las técnicas que describimos brevemente a continuación, y en especial las de prospección de datos y aprendizaje automático, aportan soluciones automatizadas que pueden ser de gran utilidad en la obtención de resultados en el análisis del consumo de información.

4.1. Prospección de datos (data mining)

La prospección de datos (*data mining*) se inscribe en un proceso más amplio como es el del descubrimiento de conocimiento dentro de grandes bases de datos (KDD o *knowledge discovery in data bases*). Este proceso, no trivial, consiste en descubrir patrones válidos en un conjunto de datos, que deben ser potencialmente útiles en relación al objetivo propuesto en el proceso de prospección de datos y comprensibles para el usuario. Molina (2002) define el *data mining* partiendo, precisamente, de la distinción entre datos, información y conocimiento. El *data mining* trabaja en un nivel superior buscando patrones de conducta, agrupaciones, secuencias, tendencias o asociaciones de datos que puedan generar algún modelo que permita entender mejor el dominio con el objetivo de facilitar la toma de decisiones.

De las fases que componen el proceso de prospección de datos, identificadas por Sangüesa (2000) nos interesa profundizar en la primera de ellas, la definición de la tarea, donde se define el objetivo a lograr en el análisis que conlleva la elección de un o más modelos de prospección a utilizar. En nuestro caso son aplicables la mayoría de dichos modelos ya que en el análisis de los datos de consumo de información nos puede interesar agrupar comportamientos similares, clasificar recursos consultados, estableciendo un ranking de uso, o analizar las habilidades informativas para predecir acciones futuras. A continuación enumeramos brevemente los diferentes modelos de prospección de datos

existentes ordenados en función del tipo de objeto preestablecido:

- Modelo de agregación (*clustering*), si nos proponemos *encontrar similitudes* y *agrupar modelos semejantes* . Un ejemplo sería localizar grupos de datos similares.
- Árboles de decisión, tanto si nuestro objetivo es *clasificar objetos* como si nos interesa obtener conocimiento para poder *hacer predicciones* .
- Redes neuronales y las reglas de clasificación, si nuestro objetivo es *clasificar objetos,* estudiar las diferencias entre grupos, sus características particulares.
- Modelos predictivos clásicos de la estadística, en el caso de que nuestro interés sea obtener conocimiento a partir de los datos que nos permita *predecir* acciones, comportamientos, etc.
- Modelos descriptivos como , las redes bayesianas y, en menor grado, las reglas de asociación, si nos proponemos encontrar y expresar asociaciones significativas o causales entre diversas variables, *hacer descripciones* .

Todos estos modelos de prospección de datos infieren los datos a partir de patrones de regularidad hecho que los hace especialmente adecuados para la definición de perfiles de usuario horizontales. A partir de nuestra hipótesis de trabajo podríamos, por ejemplo, aplicar a los datos sobre acceso a recursos Internet de los trabajadores de un organismo una técnica como el modelo de agregación (*clustering*). De este modo, podríamos obtener una relación de los recursos web consultados y establecer un ranking de los más visitados. Otro ejemplo sería aplicar a datos sobre consumo de información un modelo de prospección como son los árboles de decisión. Esta aplicación nos permitiría, con toda probabilidad, analizar el comportamiento de un usuario frente a una necesidad informativa concreta y hacer una predicción de su comportamiento en una situación similar.

4.2. Métodos olap y sistemas oltp

Los métodos OLAP (*Online Analytical Processing*) surgen de la necesidad de analizar los datos de ventas y marketing así como para procesar datos administrativos. Se alimentan de los datos generados por los sistemas transaccionales (facturación, ventas, producción, etc.). Una de las características de estos sistemas es la posibilidad de realizar un análisis multidimensional de los datos visualizados de forma tridimensional a partir de una figura cúbica. El proceso de generación de dicha representación es compleja pero la optimización de la consulta de los datos es muy alto ya que pueden ser seccionados y visualizados desde múltiples perspectivas.

Los sistemas de procesamiento de transacción en línea (OLTP) tienen como objetivo conservar la integridad de los datos necesarios para administrar una organización de manera eficiente. Los datos se presentan de forma jerárquica y dimensionada para cada empleado o tipo de empleado de la organización. Los datos se pueden visualizar desde las diferentes perspectivas definidas.

Tanto los métodos OLAP como los sistemas OLTP se basan en esquemas de trabajo apriorísticos, en los que la intervención humana es hace más presente. Por este motivo ambos por igual serian de utilidad en la definición de perfiles de usuario verticales y horizontales.

4.3. Estadística

La estadística ha estado y está dedicada al análisis de grandes volúmenes de datos. Sangüesa (2000) habla del reto que ha supuesto la prospección de datos para la estadística, debido a la necesidad de crear instrumentos de trabajo que den respuestas de fácil comprensión a usuarios no siempre expertos en su funcionamiento.

De hecho buena parte de los métodos de prospección de datos proceden de la estadística, por ejemplo, los métodos de clasificación y agregación de datos, los modelos de predicción, las redes bayesianas y los métodos heurísticos (métodos de investigación que inventan nuevos procesos, nuevas formas de organización, a partir de los que se están llevando a cabo con el objetivo de mejorarlos).

4.4. Aprendizaje automático (machine learning)

El aprendizaje automático (*machine learning*) es la rama de la inteligencia artificial que estudia el modo en que los sistemas inteligentes son capaces de desarrollar conocimiento y habilidades nuevos a partir de su experiencia. Los métodos de aprendizaje automático buscan la extracción de nuevo conocimiento a partir de la observación de los datos de su entorno o del mismo comportamiento del sistema inteligente. Este campo a aportado a la prospección de datos una gran parte de los métodos basados en la lógica, el aprendizaje basado en casos, las redes neuronales, reglas de

clasificación, etc.

Un ejemplo de aplicación de técnicas de machine learning a datos sobre consumo de información es el estudio de Chan (1999) que utiliza técnicas de aprendizaje de consumo informativo en la web para la construcción de perfiles de usuario en este entorno e introduce el concepto de indicador de interés de una página (*page interest estimator*), basado en un algoritmo de aprendizaje automático.

Para profundizar en este campo los siguientes recursos en línea destacan, especialmente, por su exhaustividad: Machine Learning Page y Online Machine Learning Resources .

4.5. Indización automática

En el apartado anterior hemos visto como los datos que más información pueden aportar sobre el consumo informativo con un grado de fiabilidad más alto son los propios contenidos de los documentos y recursos consultados. Es en este sentido que la aplicación de técnicas de indización automática es de sumo interés para el trabajo del documentalista. Concretamente queremos destacar de este ámbito de trabajo la propuesta de algoritmo de indización automática avanzada propuesta por Codina y Rovira (2000), que se basa a su vez en el modelo de Salton, y que consiste en los pasos siguientes:

- Identificación de las cadenas de caracteres, para determinar la primera lista de candidatos de términos de indización
- Eliminación de palabras vacías
- Creación de raíces de palabras derivadas con las cadenas de caracteres para crear los términos de indización
- Mezcla de términos sinónimos
- Cálculo de frecuencias absolutas
- Cálculo del peso o importancia de los términos en cada documento
- Eliminación de los términos con un índice de discriminación por debajo del umbral predeterminado
- Asignación final de descriptores ponderados a cada documento

Queremos destacar como elemento relevante en el proceso de indización automática el índice de discriminación del término. El cálculo de este índice se basa en la siguiente fórmula: la frecuencia absoluta del término en el documento por la frecuencia inversa del documento. Es decir, el número de veces que aparece el término (i) en el documento, multiplicado por el resultado de calcular el número de documentos del fondo documental partido por el número de documentos en los que aparece el término. El índice de discriminación permite dar más peso a los términos que tienen una mayor presencia en un documento y una menor incidencia en el conjunto del fondo documental.

La aplicación de estos cálculos y algoritmos son de gran utilidad en la labor de automatizar el análisis de los datos sobre consumo informativo.

5. Agentes de información web y documentales

En el proceso de estudio previo a la formulación de la propuesta de criterios para el análisis del consumo informativo ha tenido un papel muy destacado la localización de bibliografía sobre prototipos de agentes de información documentales y agentes web, que han permitido argumentar y justificar nuestra propuesta de trabajo. Por agente de información entendemos un programa que busca informaciones determinadas en Internet a partir de los criterios de búsqueda establecidos por el usuario. La documentación de experiencias y puesta en marcha de agentes de información documentales y agentes de información web nos ha sido especialmente útil dado que el objetivo final de nuestro análisis del consumo informativo es la creación de un agente de información que ofrezca a cada trabajador aquellos recursos que le son necesarios al iniciar su sesión de trabajo.

Del análisis de los agentes de información web documentados vemos que todos ellos tiene en común el hecho de basarse en el aprendizaje del comportamiento del usuario como método para inferir datos que permitan generar búsquedas paralelas a las que realiza el usuario ofreciendo resultados lo más relevante y pertinente posibles. En cambio se hace evidente una clasificación entre sistemas en los que el usuario tiene un importante peso en la evaluación de los recursos recuperados y sistemas que no prevén dicha intervención. En el primer caso encontramos agentes web como el prototipo Syskill & Webert (Pazzani, Muramatsu & Billsus, 1998), capaz de aprender el comportamiento del usuario para sugerir nuevos enlaces a partir del ranking de interés definido por el mismo, o el prototipo de los autores Balabanovic, Shoham & Yun (1995), que analiza los contenidos de las páginas consultadas a partir de métodos heurísticos de recuperación de la información, ofrece al usuario los resultados de una primera búsqueda autónoma del

agente y es el usuario el que evalúa dichos resultados.

En el segundo caso destacamos los agentes web totalmente automatizados Letizia (Lieberman, 1993), pionero en su género y cuyo objetivo principal es analizar el comportamiento del usuario y avanzarse en su búsqueda de información ejecutando consultas paralelas de manera autónoma ; y SurfLen (Xiaobin, Budzik & Hammond, 2000), con un objetivo similar al anterior y que utiliza una metodología de trabajo basada en la recogida de datos sobre la navegación del usuario y la aplicación de técnicas de prospección de datos.

En el caso de los agentes de información documentales nos han sido de interés como referencia para nuestra propuesta de análisis los trabajos de Budzik & Hammond (1999) y Ferreira & Silva (2001). Budzik & Hammond son los creadores de Watson, un asistente en la gestión de información (*information management assistant*) que incide de manera especial en la contextualización de la necesidad informativa del usuario y ofrece recursos de información justo a tiempo (*just-in-time*) a partir del análisis de los contenidos en los que está trabajando. En cambio el agente MySDI (Ferreira & Silva, 2001) es un sistema de difusión selectiva de la información personalizado que constata que el uso de técnicas de aprendizaje automático complementadas con el feedback de explícito del usuario y con la observación de su comportamiento permiten mejorar los métodos de definición de perfil de usuario usados hasta el momento.

De los ejemplos expuestos concluimos que los sistemas que automatizan completamente el proceso de análisis son, ciertamente, más ambiciosos que los que requieren de la intervención del usuario final y no por ello son menos precisos. Probablemente, el criterio humano aumenta la fiabilidad del procesos final pero también lo convierte en un sistema más definido a una casuística concreta. Los sistemas totalmente automatizados presentan, posiblemente, resultados menos pertinentes y/o relevantes al usuario, si bién la exclusión de la intervención humana los convierten en sistemas de trabajo más generalizables y extensibles a otros casos. Y en este aspecto son de interés para la elaboración de nuestra propuesta.

6. Propuesta de análisis de consumo informativo

Nuestra propuesta de trabajo la hemos dividido en cuatro fases:

- selección los datos más significativos del consumo informativo en Microsoft Office e Internet
- selección los instrumentos de localización más pertinentes en función de los datos escogidos
- definición de los parámetros a emplear en el análisis de los datos de consumo
- formulación de la propuesta de automatización de análisis de los datos a modo de algoritmo

En relación a las dos primeras fases de trabajo han sido descritas anteriormente en el punto tercero del presente artículo. Tal y como decíamos, nuestra conclusión es que los datos que más información pueden aportar sobre el consumo de información con un grado de fiabilidad más alto son los propios contenidos de los documentos y recursos consultados y sus elementos descriptivos, tanto en el caso de Microsoft Office e Internet. Todos ellos son fácilmente localizables en nuestro ordenador de trabajo. En el caso de Internet la memoria caché del programa de navegación destaca por la cantidad de información útil que recoge, incluidos los contenidos de los recursos consultados. En el caso de Microsoft Office surge como instrumento de gran utilidad la carpeta que Windows 2000 crea automáticamente y que recoge toda la actividad de un usuario concreto.

Con total seguridad la cantidad de datos será lo suficientemente elevada para justificar un método de análisis automatizado. En la tercera fase del trabajo definimos los siguientes parámetros de trabajo, ponderados de mayor a menor grado de importancia:

- Clasificación de los recursos de información consultados a partir de sus metadatos (de acuerdo con el estándar RDF, sintaxis para definir metadatos) en el casos de recursos de Internet, o de las categorías o palabras claves definidas en las propiedades de los documentos
- Asignación de las categorías resultantes a las tareas asignadas a cada usuario, que dividimos entre:
- Específicas de cada puesto de trabajo (valorando la periodicidad de dichas tareas así como la caducidad de la información necesaria)
- Generales en el funcionamiento del organismo

A nuestro entender el resultado de esta fase nos permitiría obtener una primera clasificación automatizada de los datos de consumo. Y ya en la última fase de la propuesta formulamos, de manera abstracta, el siguiente algoritmo de automatización:

- Procedimiento para los contenidos (Web y Office):
- Eliminación de las palabras vacías de contenido semántico

- Cálculo del peso representativo de los términos
- Cálculo del índice de discriminación de los términos
- Procedimiento para los datos formales de la actividad web:
- Agrupación (*clustering*) de las URL consultadas
- Agrupación (*clustering*) de las URL seleccionadas

Siguiendo estos pasos obtendríamos los primeros datos numéricos sobre el consumo de información de un grupo de usuarios. Para refinar este resultado sería oportuno analizarlo mediante un programa de aprendizaje automático (*machine learning*) que nos permitiera descubrir relaciones implícitas entre los datos. Esto implica añadir un tercer paso que es la conversión de los datos textuales a datos numéricos.

Se trata de una primera propuesta que sin duda su puesta a la práctica conllevará su redefinición y ajuste a las posibilidades técnicas que estén a nuestro abasto. Los agentes de información documentos, sin embargo, permiten argumentar las fases de trabajo definidas y asegurar la obtención de resultados con un grado de fiabilidad aceptable.

7. A modo de recapitulación

Hemos visto como nuestra actividad diaria, sea con el procesador de textos Word, sea con el programa de navegación que usemos para acceder a Internet, queda plasmada en el ordenador de trabajo usado. Son unos datos, relativamente fáciles de recuperar que nos hablan de nuestra actividad durante un período determinado de tiempo. Estos datos, que Adomavicius & Tuzhilin denominan transaccionales complementan los datos factuales a partir de los cuales se elabora, actualmente, un perfil de usuario. Es posible técnicamente conocer la actividad informativa de un usuario y extraer un comportamiento informativo concreto. Podemos saber cual es el buscador que más utiliza; deducir, a partir del registro de las palabras clave introducidas en dichas búsquedas, los temas o las necesidades informativas puntuales que ha intentado resolver; cuales son los recursos informativos más consultados a lo largo de una semana, etc.

Sin duda, y siempre dentro de los límites marcados por la ética de la información, disponer de los dato de consumo informativo de un usuario nos permite complementar la definición de perfil de usuario tal y como se ha entendido hasta ahora. La incorporación de datos transaccionales inciden en la definición de un perfil de usuario de una forma novedosa y positiva. Sin duda, la gran innovación que suponen es la posibilidad de poder establecer comparaciones entre los datos de consumo de información de un grupo de usuarios concreto. ¿Qué necesidades informativas comparten el jefe de un departamento, un analista informático y un administrativo? A simple vista no parece que haya puntos en común, pero probablemente el análisis transaccional de sus datos de consumo informativo nos ofrezcan una nueva perspectiva muy significativa desde el punto de vista de la documentación. De hecho surge como un instrumento de gran valor en la evaluación de recursos documentales, por un lado, y de definición de perfiles de usuario, del otro.

Es remarcable, a nuestro modo de ver, la interdisciplinariedad del ámbito de estudio dado que no sólo intervienen elementos propios de la recuperación de la información, sino también, como hemos visto, de la inteligencia artificial o de la prospección de datos, dando una nueva proyección a la disciplina documental.

8. Referencias bibliográficas

Adomavicius, Gediminas; Tuzhilin, Alexander.(2001) "Using data mining methods to build customer profiles ″ [en línia], dins *Computer* . [s.l.]: IEEE, pàg. 74-82. *http://citeseer.nj.nec.com/adomavicius01using.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Aha, David W. [s.d.]. Machine Learning Page [en línia]. http://www.aic.nrl.navy.mil/~aha/research/machine-learning.html [Font: Google] [Última consulta: març de 2003]

Balabanovic, Marko; Shoham, Yoav; Yun, Yeogirl (1995). "An adaptative agent for automated web browsing ″ [en línia], dins *Journal of visual communication and image representation* , *http://citeseer.nj.nec.com/balabanovic95adaptive.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Budzik, Jay; Hammond, Kristian J. [s.d.]. "User interactions with everyday applications as context for just-in-time information access ″ [en línia]. *http://citeseer.nj.nec.com/budzik00user.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Budzik, Jay; Hammond, Kristian J. (1999) "Watson: anticipating and contextualizing information needs ″ [en línia]. *http://citeseer.nj.nec.com/budzik99watson.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Chan, Philip K. "A non-invasive learning approach to building web user profiles ″ [en línia]. [Font: ResearchIndex] [Última consulta: març de 2003]

Damiani, Ernesto [et al.] (2001). "Modeling users' navigation history ″ [en línia]. [s.l.]: [s.n.]. *http://citeseer.nj.nec.com/damiani01modeling.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Davison, Brian D. (2002). "Predicting web actions from HTML content ″ [en línia]. [s.l.]: [s.n.]. *http://citeseer.nj.nec.com/518199.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Ferreira, João; Silva, Alberto (2001). "MySDI: a generic architecture to develop SDI personalised services (how to deliver the right information to the right user?) ″ [en línia]. [Lisboa]: ISEL/ INESC ID Lisboa. *http://citeseer.nj.nec.com/489601.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Hicks, David [*et al.*] (1999). "Using meta-data to support customization ″ [en línia]. [s.l.]: [s.n.]. *http://citeseer.nj.nec.com/hicks99using.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Lieberman, Henry [1995]. "Letizia: an agent that assists web browsing ″ [en línia]. [s.l.]: [s.n.]. *http://citeseer.nj.nec.com/lieberman95letizia.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Molina Félix, Luis Carlos (2002). "Data mining: torturant les dades fins que confessin ″ [en línia]. [s.l.]: Universitat Oberta de Catalunya. http://www.uoc.edu/web/cat/art/uoc/molina1102/molina1102.html [Última consulta: març de 2003]

Online Machine Learning Resources [en línia]. http://www.ai.univie.ac.at/oefai/ml/ml-resources.html [Última consulta: març de 2003]

Pazzani, Michael; Muramatsu, Jack; Billsus, Daniel (1998). "Syskill & Webert: identifying interesting web sites ″ [en línia]. [s.l.]: [s.n.]. *http://citeseer.nj.nec.com/pazzani98syskill.html* [Font: ResearchIndex] [Última consulta: març de 2003]

Sangüesa Solé, Ramón (coord.) (2000). *Data mining: una introducció* . Barcelona: Universitat Oberta de Catalunya.

Xiaobin Fu; Budzik, Jay; Hammond, Kristian J. (2000). "Mining navigation history for recommendation ″ [en línia]. [s.l.]: Infolab, Northwestern University *http://citeseer.nj.nec.com/fu00mining.html* [Font: ResearchIndex] [Última consulta: març de 2003]

9. Notas

[1] Este artículo es resultado del trabajo de final de carrera presentado en la Universitat Oberta de Catalunya con fecha junio de 2002 para la obtención de la licenciatura en Documentación.